Creating artificial human genomes using generative neural networks

Year: 2021 | Citations: 112 | Authors: Burak Yelmen, A. Decelle, L. Ongaro, Davide Marnetto, Corentin Tallec

Abstract

Generative models have shown breakthroughs in a wide spectrum of domains due to recent advancements in machine learning algorithms and increased computational power. Despite these impressive achievements, the ability of generative models to create realistic synthetic data is still under-exploited in genetics and absent from population genetics. Yet a known limitation in the field is the reduced access to many genetic databases due to concerns about violations of individual privacy, although they would provide a rich resource for data mining and integration towards advancing genetic studies. In this study, we demonstrated that deep generative adversarial networks (GANs) and restricted Boltzmann machines (RBMs) can be trained to learn the complex distributions of real genomic datasets and generate novel high-quality artificial genomes (AGs) with none to little privacy loss. We show that our generated AGs replicate characteristics of the source dataset such as allele frequencies, linkage disequilibrium, pairwise haplotype distances and population structure. Moreover, they can also inherit complex features such as signals of selection. To illustrate the promising outcomes of our method, we showed that imputation quality for low frequency alleles can be improved by data augmentation to reference panels with AGs and that the RBM latent space provides a relevant encoding of the data, hence allowing further exploration of the reference dataset and features for solving supervised tasks. Generative models and AGs have the potential to become valuable assets in genetic studies by providing a rich yet compact representation of existing genomes and high-quality, easy-access and anonymous alternatives for private databases.